Ecological Indicators 93 (2018) 387–396

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Original Articles

Modelling species habitat suitability from presence-only data using kernel density estimation

Guiming Zhang[a,b], A-Xing Zhu[c,d,e,f,b,*], Steve K. Windels[g], Cheng-Zhi Qin[f,c]

[a] Department of Geography & the Environment, University of Denver, Denver, USA
[b] Department of Geography, University of Wisconsin-Madison, Madison, USA
[c] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China
[d] Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Nanjing, China
[e] State Key Laboratory Cultivation Base of Geographical Environment Evolution, Nanjing, China
[f] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China
[g] Voyageurs National Park, National Park Service, International Falls, USA

ARTICLE INFO

Keywords:
Habitat suitability modelling and mapping
Presence-only data
Resource availability
Kernel density estimation
Ecological monitoring

ABSTRACT

We present a novel approach for modelling and mapping habitat suitability from species presence-only data that is useful for ecosystem and species monitoring. The approach models the relationship between species habitat suitability and environment conditions using probability distributions of species presence over environmental factors. Resource availability is an important issue for modelling habitat suitability from presence-only data, but it is in lack of consideration in many existing methods. Our approach accounts for resource availability by computing habitat suitability based on the ratio of species presence probability over environmental factors to background probability of environmental factors in the study area. A case study of modelling and mapping habitat suitability of the white-tailed deer (*Odocoileus virginianus*) using presence locations recorded in aerial surveys at Voyageurs National Park, Minnesota, USA was conducted to demonstrate the approach. Performance of the approach was evaluated through randomly splitting the presence locations into training data to build the model and test data to evaluate prediction accuracy of the model (repeated 100 times). Results show that the approach fit training data well (average training area under the curve AUC = 0.792, standard deviation SD = 0.029) and achieved better-than-random prediction accuracy (average test AUC = 0.664, SD = 0.025) that is comparable to the state-of-the-art MAXENT method (average training AUC = 0.784, SD = 0.021; average test AUC = 0.673, SD = 0.027). In addition, the suitability-environment responses modelled using our approach are more amenable to ecological interpretation compared to MAXENT. Compared to modelling habitat suitability purely based on species presence probability distribution (average training AUC = 0.743, SD = 0.030; average test AUC = 0.645, SD = 0.023), incorporating background distribution to account for resource availability effectively improved model performance. The proposed approach offers a flexible framework for modelling and mapping species habitat suitability from species presence-only data. The modelled species-environment responses and mapped species habitat suitability can be very useful for ecological monitoring at ecosystem or species level.

1. Introduction

Habitat suitability modelling, also referred to as environmental niche modelling or species distribution modelling (Franklin and Miller, 2009), is essential to understanding species habitat requirements and identifying drivers of species distribution (Elith and Leathwick, 2009; Graham et al., 2004b; Leathwick and Austin, 2001; Mac Nally, 2000). Habitat suitability mapping is achieved by projecting habitat suitability models from environmental space to geographic space to predict spatial variation of species habitat suitability. The resultant habitat suitability models (i.e., species-environment responses) and habitat suitability maps can be used to support a wide range of applications such as ecological monitoring, biodiversity assessment, biological reserve design, habitat restoration, invasive species management, etc. (Ferrier

* Corresponding author at: Jiangsu Center for Collaborative Innovation in Geographical Information Resource and Application, School of Geography, Nanjing Normal University, 1 Wenyuan Road, Xianlin University District, Nanjing, Jiangsu 210023, China.
E-mail address: azhu@wisc.edu (A.-X. Zhu).

https://doi.org/10.1016/j.ecolind.2018.04.002
Received 30 September 2017; Received in revised form 1 February 2018; Accepted 1 April 2018
1470-160X/

et al., 2002; Lindenmayer and Likens, 2010; Telesco et al., 2007; Thorn et al., 2009; Thuiller et al., 2005).

The key to habitat suitability modelling and mapping is deriving the relationships between species habitat suitability and environmental conditions (i.e., environmental niches) from species data and environmental data (Guisan and Zimmerman, 2000; Hirzel and Lay, 2008; Warren, 2012). With the rapid development of geospatial technologies such as geographic information system (GIS) and remote sensing, environmental data are increasingly available (Gillespie et al., 2008; Kerr and Ostrovsky, 2003; Viña et al., 2008). According to the different species data required for deriving the suitability-environment relationships, methods for habitat suitability modelling and mapping fall into three groups: presence-absence methods, presence-pseudo-absence methods, and presence-only methods (Elith and Leathwick, 2009).

Presence-absence methods require both species presence and absence data to derive suitability-environment relationships. Examples are generalized linear models (GLM), generalized additive models (GAM) (Guisan et al., 2002), and regression trees (De'Ath, 2002). Presence-absence data are often collected through well-designed biological surveys. Thus presence-absence data are usually only available for a target group of species in small geographic areas (Brooks, 2004; Pressey, 2004). Even if absence data are available, accuracy of the data can be problematic. A recorded absence might simply result from the failure to detect the species, or the habitat was suitable but not accessible for the species. In either case, a recorded absence is not a true absence of the species (Gu and Swihart, 2004; Hirzel et al., 2002; Li and Hilbert, 2008).

Presence-pseudo-absence methods use pseudo-absence data (e.g., locations randomly selected in the study area) to replace absence data to train presence-absence models (Elith and Leathwick, 2007; Engler et al., 2004). However, performance of presence-pseudo-absence methods are shown to be very sensitive to strategies used to generate the pseudo-absence data, and there is no consensus of a robust strategy for generating pseudo-absences (Chefaoui and Lobo, 2008; Hanberry et al., 2012; Stokland et al., 2011; Wisz and Guisan, 2009).

Presence-only methods require only species presence data to derive suitability-environment relationships. This group of methods are widely applied in practice as many biological datasets consist of only species presence records, such as historical collections from museums and herbaria (Graham et al., 2004a) and patrol records (Zhang et al., 2017b). Among existing presence-only methods, envelope-based BIOCLIM (Busby, 1991) and HABITAT (Walker and Cocks, 1991) treat habitat suitability as invariant at locations (in environmental space) within the environmental envelopes constrained by the outermost species presences. Thus, they tend to oversimplify the ecological reality that species habitat suitability may vary even within the environmental envelopes.

Environmental similarity-based DOMAIN (Carpenter et al., 1993), EDGM (environmental-distance geometric mean) (Hirzel and Arlettaz, 2003) and LIVES (limiting variable and environmental suitability) (Li and Hilbert, 2008) compute habitat suitability at a location based on either the maxim or the geometric mean of the environmental similarities between the location and all known presence locations. Environmental distance-based SVMs (one-class support vector machines) (Guo et al., 2005) and MDMs (Mahalanobis-distance models) (Farber and Kadmon, 2003) compute habitat suitability at a location based on the environmental distance from the location to the center of all known presence locations. Implicitly, similarity- or distance-based methods model habitat suitability as a linear function of environment similarity or distance. Thus, they tend to oversimplify the ecological reality that species habitat suitability may responds nonlinearly to environmental gradient.

GARP (genetic algorithm for rule-set production) (Stockwell, 1999) and MAXENT (maximum entropy) (Phillips et al., 2006) are machine learning algorithms capable of fitting sophisticated rules or function relations (e.g., nonlinear) based on species presence data and background data. GARP and MAXENT can often achieve high prediction accuracy in habitat suitability mapping (Elith et al., 2006). But the fitted suitability-environment relationships are often implicit, complex and hard to interpret. GARP and MAXENT are mostly used for predictive mapping of species habitat suitability but have limited power for modelling species environmental niches.

ENFA (ecological niche factor analysis) (Hirzel et al., 2002) models habitat suitability based on frequency distributions of species presence over ecological niche factors (transformed from original environmental predictors using a procedure similar to principal component analysis). ENFA can accommodate nonlinear responses of species habitat suitability to niche factors. But it assumes that the frequency distribution on each factor must be unimodal and symmetrical. It thus oversimplifies the ecological reality that frequency distribution of species presence on niche factors may be multimodal or skewed. Zhu et al. (2015) models species habitat suitability using species presence probability distributions over environmental factors without assuming the unimodality or symmetry of the distributions. However, both ENFA and Zhu et al. (2015) model habitat suitability purely based on the probability distribution of species presence over environmental factors without accounting for background distributions (i.e., probability distribution of environmental factors in the study area). Neither of them accounts for the "availability" of resources, which is an important consideration when modelling habitat suitability from presence-only data (Boyce et al., 2002; Johnson et al., 2006).

This article presents a novel approach for modelling and mapping habitat suitability from species presence-only data. The approach models suitability-environment relationships using probability distributions of species presence over environmental factors. It imposes no assumptions on the shape of species suitability-environment relationships. Moreover, it accounts for resource availability by adjusting the presence probability distributions with background probability distributions. Details of the approach are presented in Section 2. A case study of habitat suitability modelling and mapping for the white-tailed deer (*Odocoileus virginianus*) at the Voyageurs National Park to demonstrate the approach is reported in Section 3. Discussion and conclusions are presented in Section 4 and Section 5, respectively.

2. Methodology

2.1. Basic idea

2.1.1. Approximating species environmental niche

The theoretical basis of habitat suitability modelling lies on the concept of species environmental niche, which characterizes how species fitness (habitat suitability) responds to environmental conditions (Guisan and Zimmerman, 2000; Hirzel and Lay, 2008; Leibold, 1995). Probability distribution of species presence over environmental gradients is often taken as a natural approximation to species (realized) environmental niche based on which species habitat suitability is modeled. For example, ENFA computes species habitat suitability based on the frequency distribution of species presence over ecological niche factor axes (Hirzel et al., 2002). MAXENT estimates a probability density surface of species occurrence (habitat suitability) over pixels in the study area, with probability density at each pixel related to environmental conditions at that location (Phillips et al., 2006).

Our approach also uses probability distributions (probability density functions) of species presence over environmental factors to approximate species environmental niches and to compute habitat suitability. The approach estimates probability distribution from species presence-only data in a nonparametric fashion. It imposes no assumptions on the shape of the distribution (e.g., unimodality, symmetry), nor on the form of species suitability-environment relationships (e.g., linear, gaussian).